Author Names & Affiliations

- Kaushik De - University of Texas at Arlington
- Paolo Calafiura - Lawrence Berkeley Laboratory

Contact Email Address (for NSF use only)

(Hidden)

Research Domain, discipline, and sub-discipline

Physics, High Energy Physics

Title of Submission

Connected and Scalable Cyberinfrastructure for the LHC by 2030

Abstract (maximum ~200 words).

In 2030, the ATLAS experiment will be collecting data at the Large Hadron Collider (LHC) to search for new particles and forces of nature and perform measurements of the fundamental laws of physics. The volume of data collected by 2030 is expected to be a factor of 40 larger than the data collected during the recently completed Run 1 at the LHC - which will eventually grow to a factor of 100 by 2035. ATLAS processes more than an exabyte of data per year since 2013, which will grow with time. The complexity of the data and analysis techniques will also increase in the future. The architecture and capabilities of computing resources will continue to evolve. ATLAS will require a well-connected and scalable distributed cyberinfrastructure to meet the enormous computing challenges of the future. The NSF plays a key role in US national cyberinfrastructure, which is a cornerstone of the success of ATLAS physics goals - we need continuing and expanding support for it through the next two decades. New innovations in cyberinfrastructure will also be crucial to meet the daunting challenges of LHC computing.

Question 1 Research Challenge(s) (maximum ~1200 words): Describe current or emerging science or engineering research challenge(s), providing context in terms of recent research activities and standing questions in the field.

The ATLAS experiment at the Large Hadron Collider (LHC) will continue to search for new particles and forces of nature and perform measurements of the fundamental laws of physics for many decades to come. The experimental program at the LHC has already led to the Nobel Prize winning discovery of the Higgs Boson, providing crucial insight into the nature of mass. An exhaustive hunt for Dark Matter particle(s) is underway to discover the true nature of a majority of the unknown matter in the universe. Further details of the scientific goals of the ATLAS experiments is available at http://atlas.ch/.

Run 2 at the LHC is currently underway with the accelerator providing colliding particles with 62.5% more energy than Run 1. The total collected data is expected to be 5-6 times the data previously collected from Run 1. Due to the increased volume and complexity of the data, ATLAS will require additional computing capacity 5-10 times larger than Run 1, over the next few years. By 2023, Run 3 at the LHC is expected to provide another factor of 2 more data.

The high luminosity LHC, planned to be operational in 2026, would provide unique opportunities for exploration of the Higgs boson and any new phenomena discovered at the LHC. Run 4 at the HL-LHC will provide 10 times more data than Run 3 - which will be a huge scaling challenge. The higher luminosity will create on average 200 collisions at each experiment per beam crossing. Recent studies have showed that such event pileups require 25 times longer computing time to simulate and digitize collision events. In conjunction with 100 times more accumulated data compared to Run 1, computing capabilities at even higher scale may be necessary in twenty years, which would require big improvements in both software and facilities in order to be feasible. This will present both a challenge and an opportunity for NSF cyberinfrastructure. Software technologies for virtualization and containerization, storage federation, dynamic network control, and other areas important to distributed computing will be required along with distributed facilities at bigger scale.

Higher beam energies and intensities at the LHC will produce collision events that are far more complex to analyze. Current data analysis techniques will not be sufficient to extract new physics results with the expected computing resources. The ATLAS collaboration is evaluating new Machine Learning pattern recognition algorithms, and in particular deep learning neural networks, a class of algorithms that scales well with event complexity and that are well-suited to run in energy efficient, data parallel architectures such as GPGPUs, FPGAs, and neuromorphic platforms that may well be an important component of future cyberinfrastructure and algorithms designed for them.

Question 2 Cyberinfrastructure Needed to Address the Research Challenge(s) (maximum ~1200 words): Describe any limitations or absence of existing cyberinfrastructure, and/or specific technical advancements in cyberinfrastructure (e.g. advanced computing, data infrastructure, software infrastructure, applications, networking, cybersecurity), that must be addressed to accomplish the identified research challenge(s).

Traditionally, computing for the LHC experiments have been carried out at distributed high throughput grid computing facilities. The growth of these resources, assuming constant funding levels, is expected to be much slower than the growth of data volumes and data complexity at the LHC. Recently, the LHC experiments have demonstrated successful utilization of HPC resources for collision simulations at facilities like Comet, Edison, MIRA, Stampede, and Titan. Such utilization of HPC resources in conjunction with more high throughput grid and cloud computing resources will be required in the future for full realization of the scientific goals of the LHC program.

The computing model that emerged during the successful publication of over five hundred journal articles by the ATLAS experiment over the past 5-6 years is far more complex than originally envisioned. The ATLAS experiment is the largest and most complex sensor system ever built. Dozens of independently developed applications, with tens of millions of lines of legacy code, are required to analyze the data from these sensors. Thousands of physicists require access to distributed data worldwide. For example, since 2013, more than an exabyte of data is processed every year by ATLAS physicists on distributed computing systems. As the scale of processing grows, ATLAS will require hundreds of petabytes of storage at each computing center. The centers need to be connected to wide area networks with terabyte per second links in the future. Large amounts of archival storage capacity will be needed for data preservation.

At the most basic level, since we expect faster scaling up of resource needs (computing cycles, storage and networking) than possible through constant level of funding, the distributed computing model of ATLAS will have to scale up to a larger and more diverse collection of resources. While ATLAS routinely uses grid, cloud and HPC resources currently, software evolution will be required to make effective use of the complex cyberinfrastructure. All ATLAS workloads and all ATLAS users should be isolated from the heterogeneity of the resources, both computing and storage, through the PanDA portal, as is the current practice. Better integration with archival storage for data preservation will be needed. All of these goals require a complex software stack which may need revolutionary upgrades.

In summary, heterogeneous and connected resources are an important ingredient in the scalability of the LHC experiments for HL-LHC operations. Computing resources at more than 100 times the current scale are projected to be necessary for the physics program at the LHC within the next twenty years. Without a large seamless collection of diverse computing resources, and the complex software tools to enable their utilization, the ambitious physics goals of discoveries and measurements at the highest energy frontier would be severely limited, and continuing progress in our fundamental knowledge of the physical universe would become untenable.

Question 3 Other considerations (maximum ~1200 words, optional): Any other relevant aspects, such as organization, process, learning and workforce development, access, and sustainability, that need to be addressed; or any other issues that NSF should consider.

Software sustainability is becoming an important consideration as the LHC becomes more mature and evolves towards the HL-LHC. Training of students and postdocs in modern computing techniques plays an important role in sustainability and workforce development. Creating suitable career paths for data scientists and software engineers dedicated to LHC computing will be crucial to the long-term success of the LHC scientific program.

Consent Statement